
SBA
Research

SBA Research

COMET
Competence Centers for Excellent Technologies

WORKSHOP 2: ANONYMISIERUNG

27.09.2017

SBA Research

CONTINUOUS (STREAMING) DATA ANONYMIZATION

Problems and Solution Approaches

Basics

Traditional data is

* finite

* persistent

Streaming data is

* continuous

* potentially infinite

* time varying (concept drift !?)

=> Generally we want to analyze (almost) in real time

General streaming problems

* Unordered (out-of-sequence)
* Delays
* High latency
* Multiple streams
* "Skew": Delay between event & and processing time
* If a person appears more than k times, and their 'packages' arrive immediately in sequence, they could fill up an equivalence class without being generalized at all, increasing the risk of re-identification significantly

Streaming anonymization problems

* local recoding incurs less information loss, although still not mathematically optimal

  - since k-anonymization is combinatorial optimization

  - and the number of possible combinations is exponential (k-anonym. even NP hard)

  - but greedy clustering is still polynomial (O(n^2))

* determining a global generalization level incurs more information loss

  - but can be done in (near) linear time

  - phase 1: observation and determination

  - phase 2: generalizing while instances are coming in

Criteria achievable via greedy streaming

* k-anonymity: Yes, if we simply 'fill buckets'
  - can't do it one-by-one => generalization level would depend on instance arrival order
  - but we will definitely over-generalize
* l-diversity: Yes, same as k-anonymity
  - but it will compound our data loss problem by forcing us to observe the l-criterion immediately
* t-closeness: Not precisely, as we do not know the global distribution of attribute values yet

Algorithm 1: CASTLE (adaptive clustering)

* n-bucket approach
* Basic ideas:
  - sequential cluster creation (inverse to SaNGreeA)
  - guarantees on time delay (e.g. for outlier detection)
* infinite append-only sequence of tuples
* QI attributes define a metric space, such that tuples can be considered points in this space
* cluster C over input is defined as a set of intervals, called range intervals, in the quasi-identifier attribute domains

Algorithm 1: CASTLE - cont'd

* In case a new tuple cannot be added to an existing cluster, a new cluster is formed with this instance

* also, tuples cannot exceed a certain age.. if they do & a cluster has already reached size = k, the whole cluster is output as equivalence class

* else the cluster is merged with similar clusters & output

* threshold parameter ***t*** can also be used to adapt to stream distribution => well clustered tuples will force the stream into lower-generalized clusters

Algorithm 2: FAANST (numerical streaming)

1. fill up a window of samples first
2. then run k-means clustering on the initial windows
   - clusters which reach size >= k are output immediately
   - clusters which reach information loss < threshold are called 'accepted clusters' and memorized
3. fill up new cluster
   - instances that fall into accepted clusters are generalized the same way instantly
   - others: run k-means again

A third idea...

1. start computing gen-hierarchies after cold-start / starting window / last-day heuristics
2. generalize incoming instances according to that metric
3. Keep a hash-map in which new instances are recorded according to their QI's gen level
   - has all QI's in their generalized form together
   - the value to each key is an array
   - this collision list forms a natural equivalence class
4. Once a 'bucket' reaches size=k, output and remove it

SOLUTION FOR DISCRETE SAMPLES - SANGREEA DEMO W/ WEIGHTS

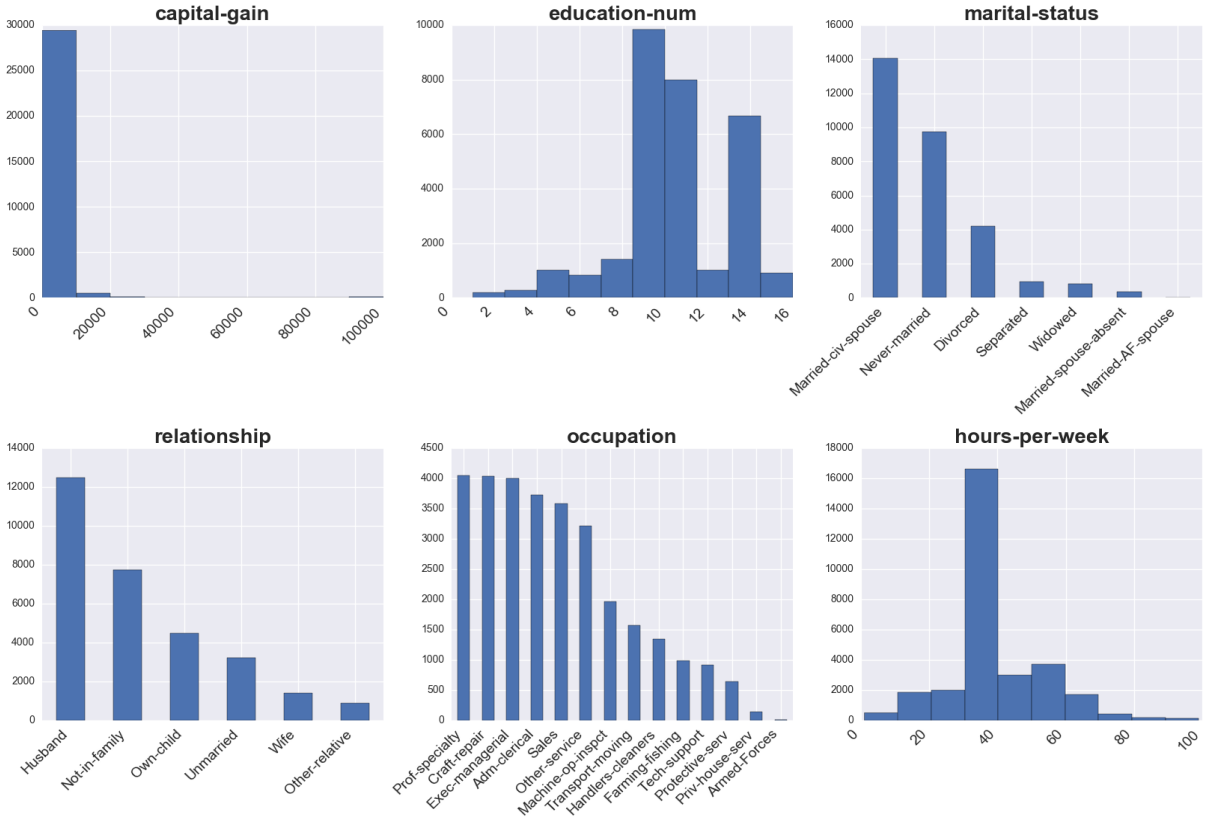
classical k-anonymization

```
## MAIN LOOP
for node in adults:
    if node in added and added[node] == True:
        continue
    # Initialize new cluster with given node
    cluster = CL.NodeCluster(node, adults, adj_list, gen_hierarchies)
    # Mark node as added
    added[node] = True
    # SaNGreeA inner loop - Find nodes that minimize costs and
    # add them to the cluster since cluster_size reaches k
    while len(cluster.getNodes()) < GLOB.K_FACTOR:
        best_cost = float('inf')
        for candidate, v in ((k, v) for (k, v) in adults.items() if k > node):
            if candidate in added and added[candidate] == True:
                continue
            cost = cluster.computeNodeCost(candidate)
            if cost < best_cost:
                best_cost = cost
                best_candidate = candidate
        cluster.addNode(best_candidate)
        added[best_candidate] = True
    # We have filled our cluster with k entries, push it to clusters
    clusters.append(cluster)
```

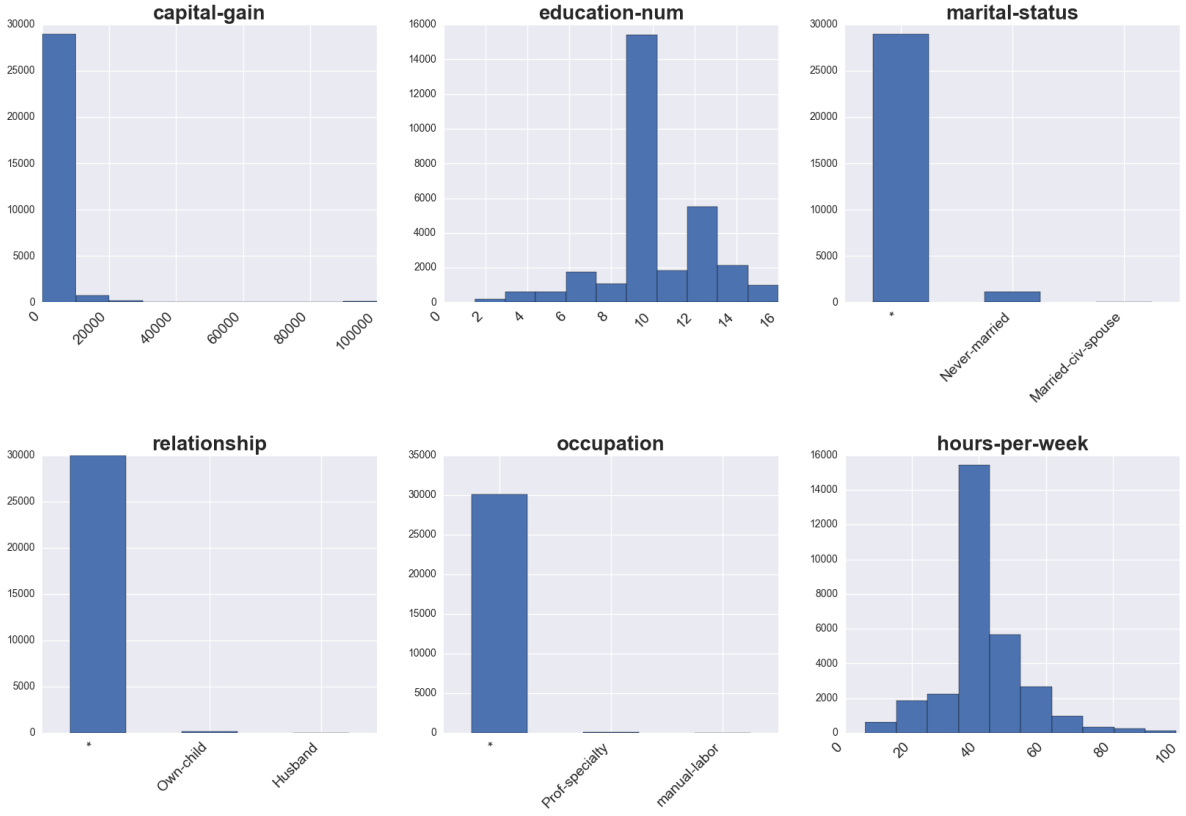
PRIVACY AWARE MACHINE LEARNING

Anonymization for a purpose

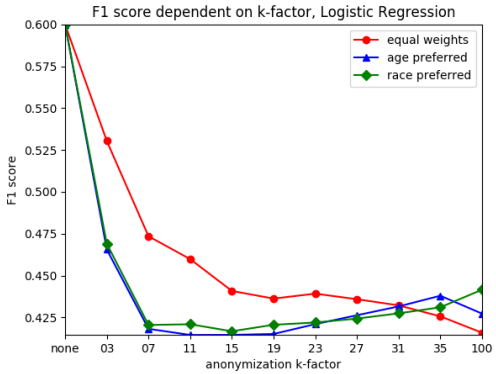
Adult data original distribution

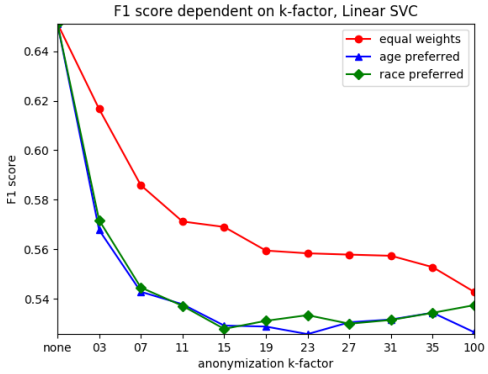


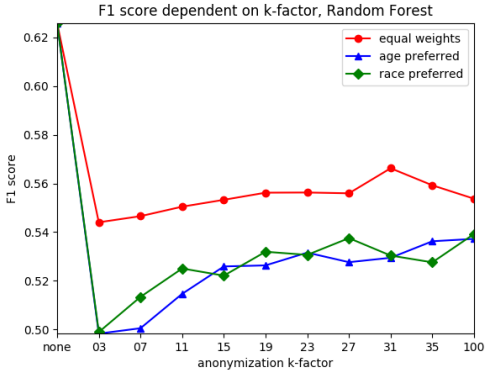
Adult data anonymized distribution

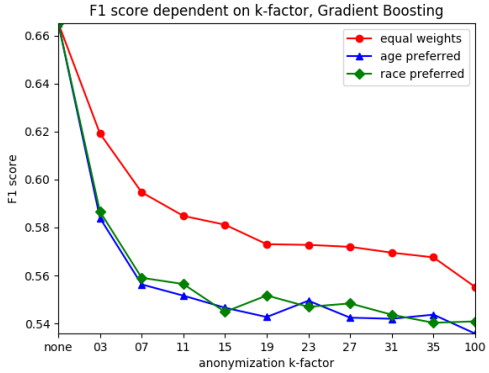


Classifier performance on education

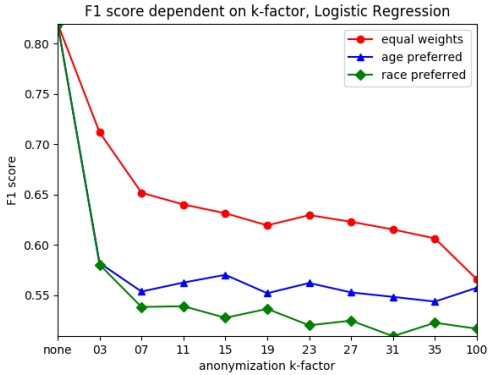


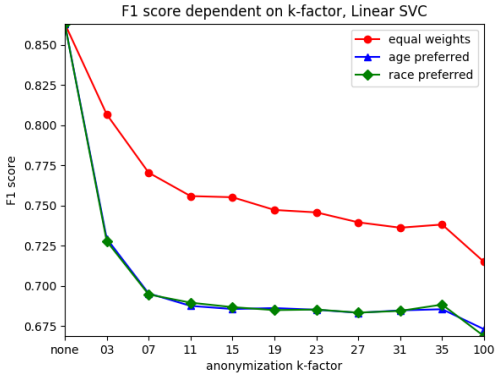


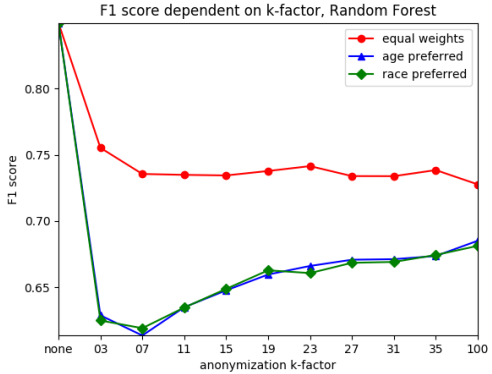


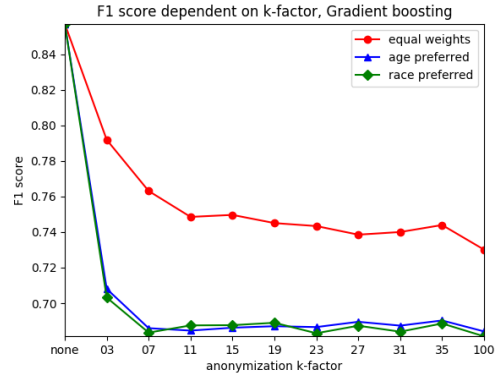


Classifier performance on marital status

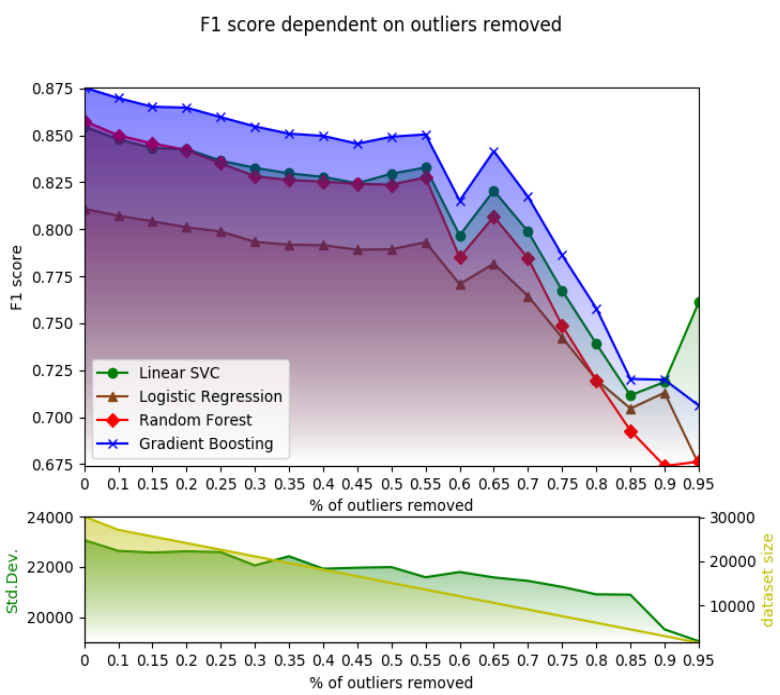


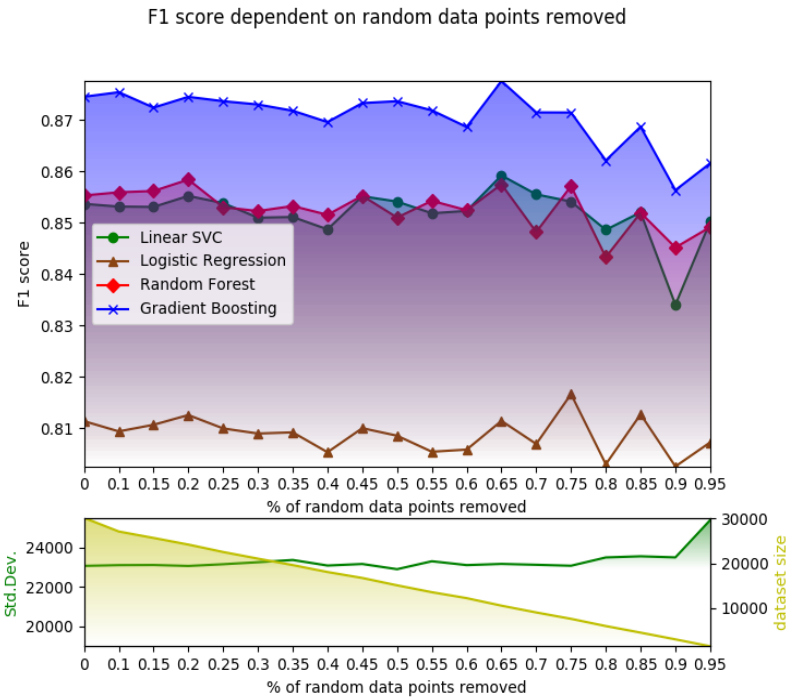




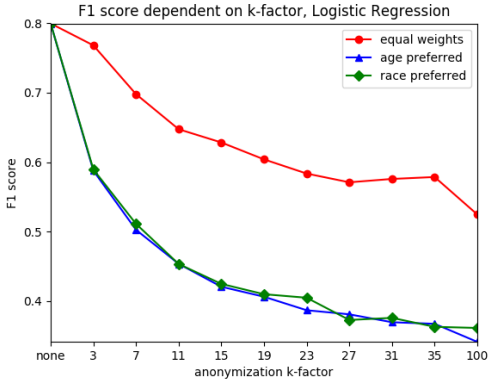


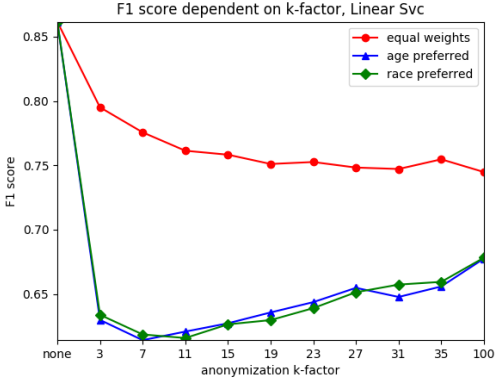
Classifier on outliers (=variance) removed

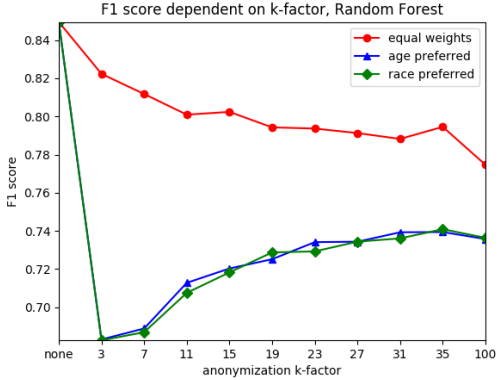


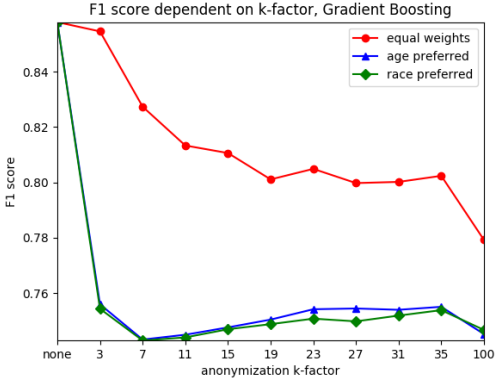


Classifiers on outliers removed => anon

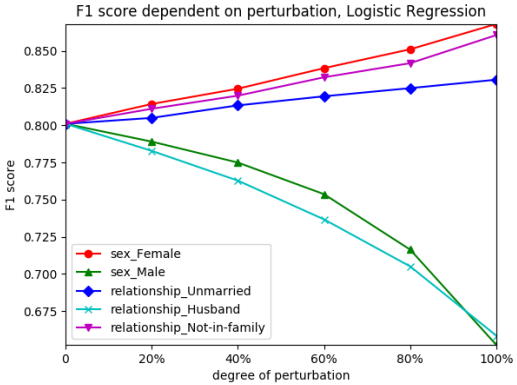


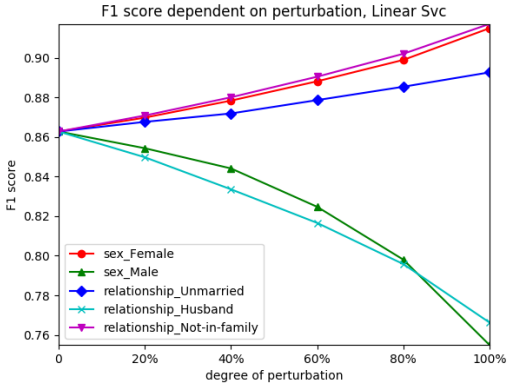


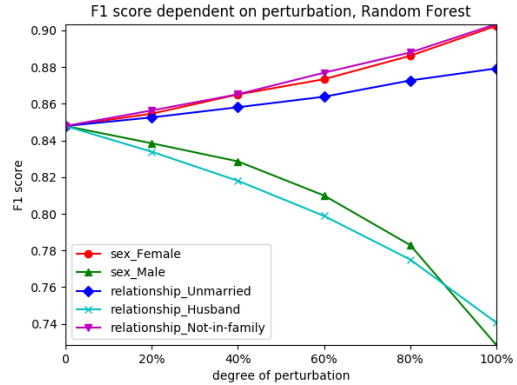


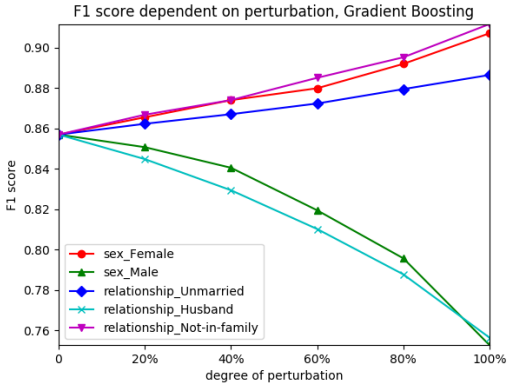


Comparison: simple data deletion









Different preferences for data preservation

Please move the data record to one cluster (up or down) with the more relevant data

age	education-num	hours-per-week	workclass	native-country	sex	race	relationship	occupation	income
[31,31]	[10,10]	[40,40]	Private	United-States	Male	*	*	manual-labor	<=50K
[31,31]	[10,10]	[40,40]	Private	United-States	Male	*	*	manual-labor	<=50K

age	education-num	hours-per-week	workclass	native-country	sex	race	relationship	occupation	income
31	10	40	Private	United-States	Male	White	Husband	Handlers-cleaners	<=50K

skip

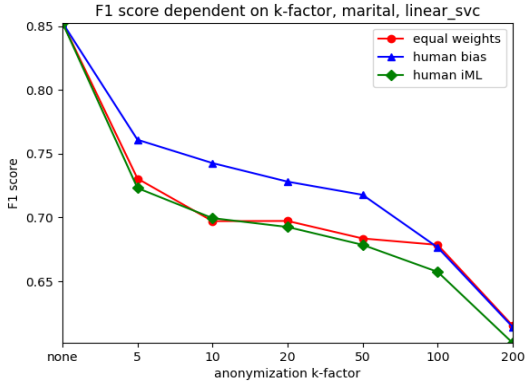
age	education-num	hours-per-week	workclass	native-country	sex	race	relationship	occupation	income
[31,31]	[10,10]	[40,40]	Private	United-States	*	White	*	*	<=50K
[31,31]	[10,10]	[40,40]	Private	United-States	*	White	*	*	<=50K

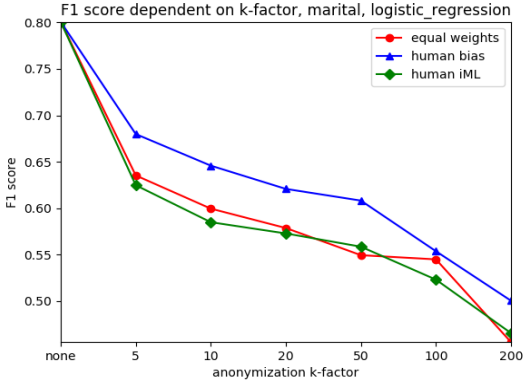
Leads to different weight vectors

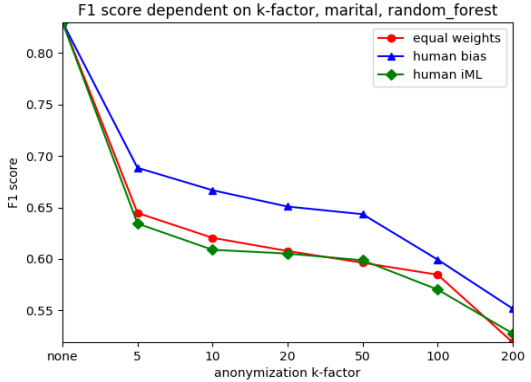
age	workclass	native-country	sex	race	marital-status
0.1667	0.1667	0.1667	0.1667	0.1667	0.1667

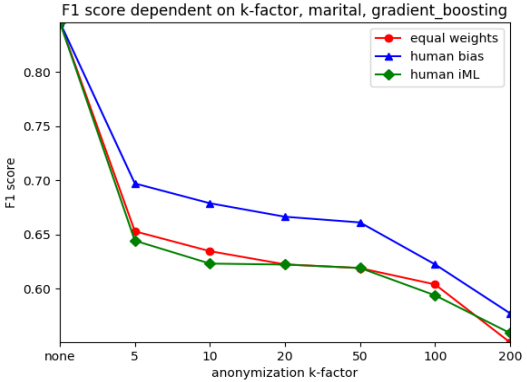
age	workclass	native-country	sex	race	marital-status
0.95	0.01	0.01	0.01	0.01	0.01

Adult data anonymized distribution









Competence Centers for Excellent Technologies

Contact: Peter Kieseberg & Bernd Malle

SBA Research gGmbH

Favoritenstraße 16, 1040 Wien

[pkieseberg, bmalle]@sba-research.org